The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/172169

Please be advised that this information was generated on 2017-09-01 and may be subject to change.
Comparative Genomic Analyses of the Moraxella catarrhalis Serosensitive and Seroresistant Lineages Demonstrate Their Independent Evolution

Joshua P. Earl1,2,3, Stefan P.W. de Vries4,5,y, Azad Ahmed3, Evan Powell3, Matthew P. Schultz3, Peter W.M. Hermans5, Darryl J. Hill6, Zhemin Zhou6, Crystala I. Constantinidou7, Fen Z. Hu1,2,3,8, Hester J. Bootsma5,9,* and Garth D. Ehrlich1,2,3,8,*

1Department of Microbiology and Immunology, Drexel University College of Medicine, Philadelphia, PA
2Center for Genomic Sciences and Center for Advanced Microbial Processing, Institute of Molecular Medicine and Infectious Disease, Drexel University College of Medicine, Philadelphia, PA
3Center for Genomic Sciences, Allegheny-Singer Research Institute, Allegheny General Hospital, Pittsburgh, PA
4Present address: Department of Veterinary Medicine, University of Cambridge, Cambridge, United Kingdom
5Laboratory of Pediatric Infectious Diseases, Radboud University Medical Centre, Nijmegen, The Netherlands
6Warwick Medical School, University of Warwick, Coventry, United Kingdom
7School of Cellular and Molecular Medicine, University of Bristol, Bristol, United Kingdom
8Department of Otolaryngology Head and Neck Surgery, Drexel University College of Medicine, Philadelphia, PA
9Present address: Centre for Infectious Diseases Research, Diagnostics and Screening, Centre for Infectious Diseases Control, National Institute of Public Health and the Environment (RIVM), Bilthoven, The Netherlands

yThese authors contributed equally to this work.

*Corresponding author: E-mail: garth.ehrlich@drexelmed.edu; hester.bootsma@rivm.nl.

Accepted: February 18, 2016

Abstract

The bacterial species Moraxella catarrhalis has been hypothesized as being composed of two distinct lineages (referred to as the seroresistant [SR] and serosensitive [SS]) with separate evolutionary histories based on several molecular typing methods, whereas 16S ribotyping has suggested an additional split within the SS lineage. Previously, we characterized whole-genome sequences of 12 SR-lineage isolates, which revealed a relatively small supragenome when compared with other opportunistic nasopharyngeal pathogens, suggestive of a relatively short evolutionary history. Here, we performed whole-genome sequencing on 18 strains from both ribotypes of the SS lineage, an additional SR strain, as well as four previously identified highly divergent strains based on multilocus sequence typing analyses. All 35 strains were subjected to a battery of comparative genomic analyses which clearly show that there are three lineages—the SR, SS, and the divergent. The SR and SS lineages are closely related, but distinct from each other based on three different methods of comparison: Allelic differences observed among core genes; possession of lineage-specific sets of core and distributed genes; and by an alignment of concatenated core sequences irrespective of gene annotation. All these methods show that the SS lineage has much longer interstrain branches than the SR lineage indicating that this lineage has likely been evolving either longer or faster than the SR lineage. There is evidence of extensive horizontal gene transfer (HGT) within both of these lineages, and to a lesser degree between them. In particular, we identified very high rates of HGT between these two lineages for β-lactamase genes. The four divergent strains are sui generis, being much more distantly related to both the SR and SS groups than these other two groups are to each other. Based on average nucleotide identities, gene content, GC content, and genome size, this group could be considered as a separate taxonomic group. The SR and SS lineages, although distinct, clearly form a single species based on multiple criteria including a large common core genome, average nucleotide identity values, GC content, and genome size. Although neither of these lineages arose from within the other based on phylogenetic analyses, the question of how and when these lineages split and then subsequently reunited in the human nasopharynx is explored.

© The Author 2016. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
**Key words:** comparative genome analysis, bacteria, Moraxella catarrhalis, evolution, phylogeny, distributed genome hypothesis, supragenome, pan genome.

**Introduction**

*Moraxella catarrhalis* is a human-restricted opportunistic respiratory tract pathogen within the pseudomonadales order of the gammaproteobacteria (www.metalife.com). This Gram-negative bacterium frequently colonizes the upper respiratory tract of young children, with two-thirds being colonized in their first year of life (Faden 2001). Although asymptomatic colonization is common, *M. catarrhalis* is an important etiological agent of upper respiratory tract infections in children; it is the third major cause of otitis media after Streptococcus pneumoniae and Haemophilus influenzae (Post et al. 1995; Aul et al. 1998; Dingman et al. 1998; Verduin et al. 2002; Murphy and Parameswaran 2009). Furthermore, it is the second most common cause of infectious exacerbations of chronic obstructive pulmonary disease, accounting for ~10% (2–4 million) of exacerbations annually in the United States (Faden 2001; Murphy et al. 2005). In rare instances it has been associated with bacteremia and septic arthritis (Melendez and Johnson 1991). The vast majority of *M. catarrhalis* strains are penicillin resistant (Hoban et al. 2001) and can form biofilms (Hall-Stoodley et al. 2006; Nistico et al. 2009) which together makes them highly recalcitrant to standard antibiotic treatment (Borriello et al. 2006; Perez et al. 2014).

Several molecular phylogenetic-typing methods, including restriction fragment length polymorphism analyses and multilocus sequence typing (MLST), have suggested that the *M. catarrhalis* species is composed of two main lineages (Bootsma et al. 2000a; Pingault et al. 2007; Wirth et al. 2007), but that divergent strains with limited homology also exist (Wirth et al. 2007). The first described phylogenetic lineage, often referred to as the seroresistant (SR) lineage, contains 16S ribotype (RB) 1 strains and displays a more pathogenic profile as the vast majority of its members are complement resistant and adhere efficiently to respiratory epithelial cells (de Vries et al. 2013, 2014). The other lineage, also known as the serosensitive (SS) lineage, consists of strains with RB2 and RB3, and mainly harbors *M. catarrhalis* isolates that are sensitive to complement-mediated killing and show less efficient adherence to respiratory epithelial cells (Bootsma et al. 2000a; Pingault et al. 2007; Wirth et al. 2007). Importantly, the two lineages differ significantly in their association with disease, with 51% of isolates from the SR lineage isolated from diseased individuals, while only 15% of the SS isolates were from diseased individuals (Verhaegh et al. 2008). Wirth et al. also demonstrated using MLST that the SS lineage is genetically more diverse than the SR lineage, and may represent remnants of an ancestral *M. catarrhalis* population. Their phylogenetic analysis indicated that the two lineages were genetically separated for an extended period. Additionally, they concluded that the genetic variation in the SS lineage appeared to be mainly the result of point mutation events, whereas SR lineage displayed increased mosaicism due to homologous recombination (Wirth et al. 2007).

Comparative whole-genome analyses of a collection of 12 clinically diverse SR isolates, each of different MLST, showed only limited genetic diversity (de Vries et al. 2010; Davie et al. 2011; Zomer et al. 2012) as compared with other nasopharyngeal pathogens including *H. influenzae* and *S. pneumoniae* (Shen et al. 2005, 2006; Hiller et al. 2007; Hogg et al. 2007; Donati et al. 2010; Boissy et al. 2011). In this study, we examined the evolution of the *M. catarrhalis* species through whole genome sequencing (WGS) of members of the SS lineage and a divergent lineage (Wirth et al. 2007) and compared these results with those obtained for the SR lineage (Davie et al. 2011).

**Materials and Methods**

*Moraxella catarrhalis* Strains and Growth Conditions

The *M. catarrhalis* SS lineage isolates and the highly divergent strains sequenced for this study (Bootsma et al. 2000a; Wirth et al. 2007; Hill et al. 2012; Stol et al. 2012) and the previously sequenced SR lineage isolates (de Vries et al. 2010; Davie et al. 2011; Zomer et al. 2012) are listed in Table 1. *Moraxella catarrhalis* strains were routinely cultured on brain heart infusion (BHI) agar plates at 37 °C in a humidified atmosphere containing 5% CO2, or in BHI broth at 37 °C in an atmosphere containing 5% CO2 with agitation (200–250 rpm). All strains including the divergent strains (Vanechoutte et al. 1990) were originally identified at the time of their isolation or characterization as Gram-negative diplococci, DNAse positive, oxidase positive, 4-methylumbelliferyl butyrate (Vanechoutte et al. 1988) positive, and glucose negative. For DNA isolations, strains were cultured until early exponential phase growth (OD620nm [optical density] of 0.2–0.4); chromosomal DNA was isolated using Genomic-tip 20G columns (Qiagen, Venlo, Netherlands) according to manufacturer’s instructions or as described (Marshall et al. 1997).

**Genome Sequencing Assembly and Annotation**

WGS was performed for 15 SS strains as described (Hiller et al. 2007; Hogg et al. 2007; Donati et al. 2010; Boissy et al. 2011; Davie et al. 2011; Ahmed et al. 2012) using a 454 LifeSciences FLX+ genome sequencer (Roche, Branford, CT) with Titanium chemistry. Two additional SS strains were sequenced using Illumina HiSeq and one SS strain was sequenced on a Pacific Biosciences RSII system using SMRT C1 chemistry according to the manufacturer’s instructions (Menlo Park, CA). The four
Table 1

*Moraxella catarrhalis* Serosensitive and Seroresistant Lineage Isolates: Strain and Molecular-Typing Data

<table>
<thead>
<tr>
<th>Strain</th>
<th>Source</th>
<th>Location</th>
<th>Reference(s)</th>
<th>Accession No.</th>
<th>Lineage</th>
<th>Ribotype</th>
<th>MLST Type</th>
<th>abcZ</th>
<th>adek</th>
<th>efp</th>
<th>fumC</th>
<th>glyRS</th>
<th>mutY</th>
<th>ppa</th>
<th>trpE</th>
<th>β-Lactamase</th>
<th>LOS Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>A9</td>
<td>NA</td>
<td>Africa</td>
<td>Bootsma et al. (2000)</td>
<td>SAMN04122787</td>
<td>SeroS</td>
<td>2</td>
<td>New 38</td>
<td>41</td>
<td>9</td>
<td>new</td>
<td>64</td>
<td>new</td>
<td>new</td>
<td>1</td>
<td>new</td>
<td>bro-1</td>
<td>C</td>
</tr>
<tr>
<td>C031</td>
<td>Nose, OM, child</td>
<td>Netherlands</td>
<td>Stol et al. (2012)</td>
<td>SAMN04122788</td>
<td>SeroS</td>
<td>2</td>
<td>ST1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>bro-1</td>
<td>A</td>
</tr>
<tr>
<td>C10</td>
<td>Sputum, COPD patient, adult</td>
<td>Netherlands</td>
<td>This study</td>
<td>SAMN04122789</td>
<td>SeroS</td>
<td>2</td>
<td>NP-ST</td>
<td>10</td>
<td>11</td>
<td>8</td>
<td>12</td>
<td>new</td>
<td>11</td>
<td>new</td>
<td>1</td>
<td>None</td>
<td>A</td>
</tr>
</tbody>
</table>

**NOTE.**—OM = otitis media; COPD = chronic obstructive pulmonary disease; LRT = lower respiratory tract; OME = chronic otitis media with effusion; NA = not available; SeroS = sero-sensitive; SeroR = Sero-resistant; Div = divergent; ND = not determined; New* = a novel MLST based on novel alleles will need to be assigned; new* = a novel ST allele based on WGS of the strain will need to be assigned; *currently the *Moraxella catarrhalis* MLST database does not accept sequence data from non-Illumina sequenced strains, thus it was not possible to get a novel ST assigned in these cases; NT = nontypeable.
divergent strains identified by Wirth et al. (2007) were sequenced with an Illumina MiSeq (Illumina, San Diego, CA). Following sequencing, the number of reads, the average read lengths, the depth of coverage, the GC content, and the number of contigs for each strain was determined. Sequencing with the 454 and Pacific Biosciences RS systems was conducted at the Center for Genomic Sciences (CGS) at the Allegheny-Singer Research Institute. As previously described (Davie et al. 2011; Ahmed et al. 2012), the 454 raw sequence reads for each strain were assembled into contigs using the Roche/454 Life Sciences’ GS de novo Newbler assembler (version 2.0.00.20) using the default parameters except for minimum overlap identity which was adjusted to obtain the fewest contigs. For all 454-based assemblies, we called single nucleotide polymorphisms (SNPs) against the self-assembly. The C10 strain, sequenced on the RS II system, was assembled using the Celera WGS-assembler (Myers et al. 2000). Illumina-sequenced strains were assembled with either Newbler (v2.6) or Ray (v2.3.1). The final genome assemblies have been deposited in GenBank, and the corresponding accession numbers are provided in table 1. The assembled genomes were submitted to RAST (Rapid Annotation using Subsystems Technology; http://RAST.nmpdr.org) for automated annotation (Aziz et al. 2008). Specific gene functions were inferred to be present in a genome if any region of that genome was inferred to be homologous to a region in another genome (see below) with the annotated function (Hogg et al. 2007).

Gene Clustering and Comparative Genome Analyses

The CGS comparative genomics pipeline (Hiller et al. 2007; Hogg et al. 2007; Donati et al. 2010; Boissy et al. 2011; Davie et al. 2011; Ahmed et al. 2012) was used to take the RAST-tfasty in the Fasta v3.6 package (Pearson and Lipman 1988). In alignments between NA and AA were done with the program acid (NA), and amino acid (AA) sequences. All-against-all strain were collated into two multi-fasta files, both as nucleic acid sequences to at least one other gene/sequence (annotated/unannotated) in the cluster (single-linkage) (Hogg et al. 2007). The gene clusters were then binned as either core or unannotated) in the cluster (single-linkage) (Hogg et al. 2007; Donati et al. 2010; Boissy et al. 2011; Davie et al. 2011; Ahmed et al. 2012). Gene Clustering and Comparative Genome Analyses

Whole-genome alignments of all 35 M. catarrhalis genomes (13 SR, 18 SS, and 4 divergent M. catarrhalis isolates), as well as an Moraxella macacae genome used as a phylogenetic outgroup were performed using default parameters in Mauve v2.3.1 (Darling et al. 2004). We also performed analyses of the SR and SS lineages using the divergent M. catarrhalis strains as an outgroup. For each genome, its locally collinear blocks (LCBs), as identified by Mauve, were concatenated together and referred to as the “concatenated core” sequence, which is independent of gene annotations. These sequences were aligned with Mafft v7.127b with default “fast” parameters. A network was inferred by importing the alignment directly into SplitsTree4 (Huson and Bryant 2006), which automatically calculates the network. Additionally, this alignment was also used as input to the program RAxML v8.0.20 (Stamatakis 2006) wherein the GTGRAMMA model was used and one thousand bootstraps were performed to measure tree branching consistency as follows: raxhpic-pthreads5-sse3 -f a -s input.fasta –n boot –m GTRGAMMA --d 1234 --x 1234 --#1000 --T 16 –n T20 –o output

Genomic Recombination Analyses

Moraxella catarrhalis is a naturally competent species. Although it is possible to explain reticulation within the phylogenetic networks as indicative of recombination events (Huson and Bryant 2006), we also explored an additional analysis to provide more rigorous evidence of horizontal gene transfer (HGT). To investigate the number of regions within these genomes potentially undergoing HGT, the recently published orderedPainting method (Yahara et al. 2014) was used. This method infers “hotspot” regions from a multiple genome alignment by examining SNPs. All SNPs containing no gap or ambiguous characters were extracted from the concatenated core genome alignment (as described above) by a custom Ruby script and analyzed on our cluster using the orderedPainting software. From these data, we identified the recombination hotspots from which we built phylm trees. First, we used JmodelTest V 2.1.6 to calculate likelihood using all 88 possible different DNA model combinations. Using the AIC criteria, the best scoring model was found to be GTR + I + G. Phyml version 20120412 was then used to calculate the maximum-likelihood tree using the command line generated by the jModelTest program: pyml --i=align.phy --d nt --n 1 --b 0 --run_id GTG+I+G --m 012345 --f m --v e --c 4 --a e --no_memory_check --o tlr --s NNI

Strain Groupings

The allelic and gene possession neighbor grouping networks were calculated from distance matrices as described (Hall et al. 2007).
2010). Briefly, allelic pairwise distance was calculated as one minus the average percent identity of all aligned core genes, and the gene possession pairwise distance was calculated as one minus the total number of distributed genes either both present or both absent in each genome, divided by the total number of distributed genes in the supragenome. Networks were generated for these analyses in SplitsTree by importing the distance matrices which automatically computed the network image (Huson and Bryant 2006).

### Supragenome Modeling

Modeling of the lineage-specific and whole species supragenomes was performed as described using the modified Finite Supragenome Model (FSGM; Hogg et al. 2007; Boissy et al. 2011).

### Age Inference Modeling

We used BEAST v1.75, a Bayesian-based model utilizing Markov chain Monte Carlo methods to calculate the age of evolutionary divergence between the SR and SS *M. catarrhalis* lineages (Drummond and Rambaut 2007). We used the concatenated core sequence alignment of *M. macacae* and *M. catarrhalis* as input and the generally accepted date of 23.5 Ma for the cercopitoid-hominoid split as the Bayesian prior for the date of the split at the tree root (Pickford and Andrews 1981) as we reasoned that the independent evolution of these two species would have tracked with the divergence of their host ancestors. A strict fixed clock model was required for age estimates, and the general time reversible (GTR) model was used, as from the limited models available in BEAST, it had the lowest AIC/BIC (Akaike information criterion/Bayesian information criterion) score as calculated from the alignment in the program Topali v2.5.

### Phentype–Genotype Association Mapping

The web-based tool PhenoLink (Bayjanov et al. 2012) was used to link observed serum resistance/sensitivity and adherence phenotypes to genome content. To this end, strains were divided into two classes for each examined phenotype as follows: Serum resistant (>10% survival) versus serum sensitive (<10% survival); intermediate to high adherent (>20%) versus low adherent (<20%). Subsequent association analyses were performed using default parameter settings.

### In Silico Phylogenetic Analysis

The 16S ribosomal RNA (rRNA) type for the SS and SR strains was determined according to Bootsma et al. (2000a). Allelic sequences of the eight MLST genes were analyzed via [http://mlst.warwick.ac.uk/mlst/dbs/Mcatarrhalis](http://mlst.warwick.ac.uk/mlst/dbs/Mcatarrhalis) as detailed (Wirth et al. 2007). Lipooligosaccharide (LOS) serotyping was done according to Edwards et al. (2005), and typing of the beta-lactamase *bro* gene was performed according to Bootsma et al. (1996). An attempt was made to deposit the MLST data, but the curator informed us the site is down for reconfiguration.

### Serum Resistance Assay

*Moraxella catarrhalis* isolates were cultured to mid-log phase (OD$_{620\text{nm}}$ ~ 1.0). Cultures were washed in four volumes of phosphate buffered saline (PBS) supplemented with 0.15% gelatin (PBS-G) and resuspended in five volumes PBS-G supplemented with 1 mM CaCl$_2$ and 0.2 mM MgCl$_2$ (PBS-G Mg$^{2+}$/Ca$^{2+}$). Resuspended bacterial cultures were mixed with an equal volume of 80% pooled normal human serum (NHS) (GTI diagnostics) diluted in PBS-G Mg$^{2+}$/Ca$^{2+}$ or NHS that was heat inactivated for 30 min at 56°C (NHS-HI). Colony-forming units (CFUs) were determined at 0 and 60 min postincubation by plating 10-fold serial dilutions on BHI plates. The survival percentage was calculated relative to NHS-HI (n ≥ 3) and statistical significance was determined using a Mann–Whitney test in GraphPad Prism 5.0.

### Adhesion to Respiratory Tract Epithelial Cells

Adhesion of *M. catarrhalis* strains to the human pharyngeal epithelial cell line Detroit 562 (ATCC CCL-138) and the type II alveolar epithelial cell line A549 (ATCC CCL-185) was performed as detailed in de Vries et al. (2009). Briefly, cells were seeded in 24-well plates at 2 × 10$^5$ Detroit 562 cells/well 2 days prior to the experiment (at 24 h the medium was refreshed), or 4 × 10$^5$ A549 cells/well 1 day prior to the experiment. For both cell lines, monolayers of approximately 1 × 10$^6$ cells/well were used for adherence assays. *Moraxella catarrhalis* isolates were cultured to mid-log phase (OD$_{620\text{nm}}$ ~ 1.0) and stored in the presence of 20% glycerol at −80°C. Cells were washed twice with PBS before the bacteria were allowed to adhere to the epithelial cells (multiplicity of infection, 10:1) for a 1 h period in infection medium (DMEM-GlutamaxTM-I with 1% fetal calf serum). Thereafter, cells were washed three times with PBS to remove nonadherent bacteria. After detachment and lysis of the epithelial cells through addition of 1% saponin (Sigma-Aldrich, St. Louis, MO) in PBS-G, CFUs were enumerated by plating 10-fold serial dilutions on BHI plates. The percentage adherence (n ≥ 4) was calculated as the percentage of the inoculum that bound to epithelial cells. Statistical significance was determined using a Mann–Whitney test in GraphPad Prism 5.0.

### Results

#### Whole-Genome Sequencing of *Moraxella catarrhalis* Strains

Previous gene-based phylogenetic studies have suggested that the *M. catarrhalis* species contains at least two distinct groups, with the SR lineage strains forming a unique clade separate from the larger SS grouping(s). Phenotypically, the SR and SS
lineages have been shown to differ in their capacity to withstand the action of the human complement system and to adhere to host epithelial cell lines (Bootsma et al. 2000a; Pingault et al. 2007; Wirth et al. 2007; Verhaegh et al. 2008). In this study, using WGS and comparative genome analytics, we endeavored to determine the total clade structure of *M. catarrhalis* and any ancestral relationships among the various subpopulations identified. Previously, work by Davie et al. (2011) described the sequencing and comparative genomic analyses of 12 SR lineage strains, all of 16S ribotype 1 (RB1), which were selected on the basis of their divergent clinical and geographic sites of isolation. To gain further insight into the overall genetic diversity of the *M. catarrhalis* species as a whole, in this study 18 SS lineage strains, one additional SR lineage strain, and four divergent strains (which by MLST typing did not cluster with either the SR or SS lineages; Wirth et al. 2007) were subjected to whole-genome sequence analysis (table 1). The SS isolates were selected based on their genetic diversity, as determined both by 16S ribotypes (RB2 (n = 12) and RB3 (n = 6)) and MLST data (Bootsma et al. 2000; Pingault et al. 2007; Wirth et al. 2007), as well as for their clinical sites of recovery, which represented diverse niches from both diseased and healthy hosts. The sequencing platforms, assembly softwares, levels of coverage, numbers of contigs, genome size, and GC content for all 35 *M. catarrhalis* strains are presented in summary form (table 2).

The strain-specific genomic data were then analyzed comparatively to obtain average measures of diversity within and among the three groups of strains. This was accomplished by computing the average of both the percent GC base content and genome size (table 2), as well as determining the average nucleotide identity (ANI) (Konstantinidis et al. 2006) and tetrnucleotide identity (Richter and Rossella-Mora 2009) using both BLAST and MUMmer algorithms. Between strain pairs was 95.78% with a range of 95.6–95.91% (supplementary fig. S1, Supplementary Material online). For the SR and SS lineages, the FSGM, based on the actual number of strains (supplementary table S3, Supplementary Material online). There is, however, as discussed below, evidence of frequent HGT and moderate scale inversions, insertions, and deletions among all SS and SR strains.

We then calculated empirical values for the *M. catarrhalis* lineage-specific core genomes, distributed genomes, and suprapan genomes, as well as the species-wide values (both including and excluding the divergent strains; table 3 and fig. 1). Adding the four divergent strains to the combined core of the SS and SR lineages significantly reduced the size of the core genome from 2,318 to 2,041 genes, a >10% decrease. The observed supragenomes for the individual SR and SS lineages contained 3,240 and 3,524 genes, respectively, whereas the combined supragenome for these two lineages contained 3,854 genes. A massive increase in supragenome size to 5,684 genes was observed when the 4 divergent strains were included in the analysis. The number of distributed genes increased by 2,107 (from 1,536 to 3,643) with the addition of just four divergent strains to the combined SRSS lineage supragenome. Interestingly, the vast majority of this increase (1,905 genes) occurred with the addition of the first divergent strain which resulted in more than doubling the total number of distributed genes in the supragenome with the addition of the next three divergent genomes only adding an additional 202 distributed genes. Collectively, these data are supportive of the divergent strains being *sui generis*.

These gene content data were then used as inputs into the revised FSGM (Hogg et al. 2007; Boissy et al. 2011) to model the total diversity present in each of the SR and SS lineages, the two lineages combined, and the species as a whole including the divergent strains. To compare similar numbers of strains for each of these groupings, we made predictions based on 50 strains (supplementary table S3, Supplementary Material online). For the SR and SS lineages, the FSGM predicts that the lineage-specific supragenomes are greater than 90% complete following the addition of 13 and 11 genomes, respectively, and the species-level (excluding the divergent strains) supragenome is 90% complete following the addition of 12 genomes. The FSGM, based on the actual number of genomes sequenced (N’s = 13, 18, and 31, for the SR, SS, and species as a whole, respectively), gave supragenome size predictions of 3,240, 3,524, and 3,854 genes which are identical to the observed values. The only substantive difference

In contrast, the divergent strains have a mean genome size of 2.16 Mb (again with a range of ~100 kb), making them 13% larger on average than the combined SS and SR strain set, and even larger than the *M. macacae* genome of 2.08 Mb. *Moraxella macacae* (Ladner et al. 2013), which we used as an outgroup for the phylogenetic tree, differed from the SS, SR, and divergent groups with an average ANI of 69.83% with a range of 69.00–70.31% while its genome had a third unique GC content of 39%. A MAUVE alignment of all 31 SS and SR strains and the *M. macacae* genome revealed that the overall genome structure across the genus is reasonably conserved (supplementary fig. S1, Supplementary Material online).
between the empiric and predicted supragenome sizes was when the divergent strains were added to the combined SS and SR strain lineages, the FSGM predicted 5,373, but the observed was 5,684 indicating that the model is likely underpowered with input from only 4 divergent strains. Combined these data strongly suggest that we have essentially captured the entire supragenome(s) of the species and its component SS and SR lineages, with the probable exception of distributed genes within the divergent group.

We next performed exhaustive pairwise comparisons in which each of the 35 SS, SR, and divergent strains’ gene content was compared against every other strain’s gene content, resulting in a total of 595 such two-way gene by gene comparisons (supplementary table S4, Supplementary Material online). As expected, the mean gene possession difference values and standard deviations were least within the SR ($\overline{x}$ = 232 ± 55) and SS ($\overline{x}$ = 322 ± 87) clades, with the SR clade showing less divergence than the SS clade. When the two

### Table 2

<table>
<thead>
<tr>
<th>Strain</th>
<th>Bnum</th>
<th>Size (MB)</th>
<th>Coverage</th>
<th>Contigs</th>
<th>N50</th>
<th>% GC</th>
<th>Sequencing Method</th>
<th>Assembly Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serosensitive lineage strains</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A9</td>
<td>B743</td>
<td>1.91</td>
<td>27.4</td>
<td>24</td>
<td>220795</td>
<td>41.70</td>
<td>454 Newbler v. 2.6</td>
<td></td>
</tr>
<tr>
<td>C031</td>
<td>B620</td>
<td>1.95</td>
<td>31.9</td>
<td>47</td>
<td>122961</td>
<td>41.50</td>
<td>454 Newbler v. 2.6</td>
<td></td>
</tr>
<tr>
<td>C10</td>
<td>B621</td>
<td>1.98</td>
<td>28.6</td>
<td>$^9$</td>
<td>412591</td>
<td>41.40</td>
<td>PacBio C1 Celeria v. 7.0</td>
<td></td>
</tr>
<tr>
<td>F18</td>
<td>B744</td>
<td>1.95</td>
<td>24.3</td>
<td>26</td>
<td>131591</td>
<td>41.70</td>
<td>454 Newbler v. 2.6</td>
<td></td>
</tr>
<tr>
<td>F20</td>
<td>B745</td>
<td>1.95</td>
<td>24.4</td>
<td>61</td>
<td>89759</td>
<td>41.60</td>
<td>454 Newbler v. 2.6</td>
<td></td>
</tr>
<tr>
<td>F21</td>
<td>B618</td>
<td>1.95</td>
<td>15.7</td>
<td>52</td>
<td>106722</td>
<td>41.50</td>
<td>454 Newbler v. 2.6</td>
<td></td>
</tr>
<tr>
<td>F23</td>
<td>B746</td>
<td>1.97</td>
<td>20.5</td>
<td>56</td>
<td>125588</td>
<td>41.50</td>
<td>454 Newbler v. 2.6</td>
<td></td>
</tr>
<tr>
<td>F24</td>
<td>B747</td>
<td>1.94</td>
<td>18.9</td>
<td>40</td>
<td>182702</td>
<td>41.70</td>
<td>454 Newbler v. 2.6</td>
<td></td>
</tr>
<tr>
<td>N1</td>
<td>B748</td>
<td>1.88</td>
<td>34.1</td>
<td>22</td>
<td>231375</td>
<td>41.60</td>
<td>454 Newbler v. 2.6</td>
<td></td>
</tr>
<tr>
<td>N12</td>
<td>B617</td>
<td>1.89</td>
<td>36.1</td>
<td>26</td>
<td>195908</td>
<td>41.60</td>
<td>454 Newbler v. 2.6</td>
<td></td>
</tr>
<tr>
<td>N15</td>
<td>B622</td>
<td>1.93</td>
<td>26.4</td>
<td>45</td>
<td>190441</td>
<td>41.40</td>
<td>454 Newbler v. 2.6</td>
<td></td>
</tr>
<tr>
<td>N19</td>
<td>B749</td>
<td>1.94</td>
<td>17.2</td>
<td>71</td>
<td>80324</td>
<td>41.50</td>
<td>454 Newbler v. 2.6</td>
<td></td>
</tr>
<tr>
<td>N4</td>
<td>B615</td>
<td>1.94</td>
<td>38.4</td>
<td>46</td>
<td>186739</td>
<td>41.40</td>
<td>454 Newbler v. 2.6</td>
<td></td>
</tr>
<tr>
<td>N6</td>
<td>B616</td>
<td>1.95</td>
<td>25.2</td>
<td>40</td>
<td>190206</td>
<td>41.50</td>
<td>454 Newbler v. 2.6</td>
<td></td>
</tr>
<tr>
<td>R4</td>
<td>B619</td>
<td>1.95</td>
<td>21.3</td>
<td>41</td>
<td>153844</td>
<td>41.60</td>
<td>454 Newbler v. 2.6</td>
<td></td>
</tr>
<tr>
<td>Z18</td>
<td>B750</td>
<td>1.88</td>
<td>17.3</td>
<td>30</td>
<td>126434</td>
<td>41.60</td>
<td>454 Newbler v. 2.6</td>
<td></td>
</tr>
<tr>
<td>MX1</td>
<td>212</td>
<td>1.9</td>
<td>449.7</td>
<td>189</td>
<td>16796</td>
<td>41.70</td>
<td>Illumina HiSeq Newbler v. 2.6</td>
<td></td>
</tr>
<tr>
<td>S11</td>
<td>213</td>
<td>1.88</td>
<td>505.5</td>
<td>120</td>
<td>30531</td>
<td>41.50</td>
<td>Illumina HiSeq Newbler v. 2.6</td>
<td></td>
</tr>
<tr>
<td>Serosensitive lineage averages</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.93</td>
</tr>
<tr>
<td>Seroresistant lineage strains</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BBH18</td>
<td>BBH18</td>
<td>1.86</td>
<td>70</td>
<td>1</td>
<td>Na</td>
<td>41.70</td>
<td>454, Illumina Newbler v. 2.6</td>
<td></td>
</tr>
<tr>
<td>7169</td>
<td>7169</td>
<td>1.9</td>
<td>182.3</td>
<td>35</td>
<td>176884</td>
<td>41.70</td>
<td>454 Newbler v. 2.3</td>
<td></td>
</tr>
<tr>
<td>103P14B1</td>
<td>103P14B1</td>
<td>1.96</td>
<td>33.4</td>
<td>99</td>
<td>74903</td>
<td>41.50</td>
<td>454 Newbler v. 2.0.01.14</td>
<td></td>
</tr>
<tr>
<td>12P80B1</td>
<td>12P80B1</td>
<td>1.81</td>
<td>23.3</td>
<td>53</td>
<td>73658</td>
<td>41.70</td>
<td>454 Newbler v. 2.0.01.14</td>
<td></td>
</tr>
<tr>
<td>46P47B1</td>
<td>46P47B1</td>
<td>1.85</td>
<td>20.6</td>
<td>69</td>
<td>53508</td>
<td>41.60</td>
<td>454 Newbler v. 2.0.01.14</td>
<td></td>
</tr>
<tr>
<td>O35E</td>
<td>B583</td>
<td>1.86</td>
<td>56.9</td>
<td>42</td>
<td>10658</td>
<td>41.70</td>
<td>Newbler v. 2.3</td>
<td></td>
</tr>
<tr>
<td>BC1</td>
<td>B496</td>
<td>1.95</td>
<td>40.3</td>
<td>43</td>
<td>143587</td>
<td>41.40</td>
<td>Newbler v. 2.0.01.14</td>
<td></td>
</tr>
<tr>
<td>BC7</td>
<td>B502</td>
<td>1.9</td>
<td>28.7</td>
<td>37</td>
<td>152590</td>
<td>41.50</td>
<td>Newbler v. 2.0.01.14</td>
<td></td>
</tr>
<tr>
<td>BC8</td>
<td>B503</td>
<td>1.91</td>
<td>43.8</td>
<td>32</td>
<td>173870</td>
<td>41.60</td>
<td>Newbler v. 2.0.01.14</td>
<td></td>
</tr>
<tr>
<td>C072</td>
<td>B507</td>
<td>1.95</td>
<td>37</td>
<td>25</td>
<td>153832</td>
<td>41.40</td>
<td>Newbler v. 2.0.01.14</td>
<td></td>
</tr>
<tr>
<td>101P30B1</td>
<td>B508</td>
<td>1.86</td>
<td>33.2</td>
<td>26</td>
<td>215599</td>
<td>41.70</td>
<td>Newbler v. 2.0.01.14</td>
<td></td>
</tr>
<tr>
<td>RH4</td>
<td>RH4</td>
<td>1.84</td>
<td>200</td>
<td>9</td>
<td>284836</td>
<td>41.60</td>
<td>Illumina Assembler v. 2.0.0</td>
<td></td>
</tr>
<tr>
<td>43617</td>
<td>43617</td>
<td>1.91</td>
<td>NP*</td>
<td>41</td>
<td>89047</td>
<td>41.70</td>
<td>NP*</td>
<td></td>
</tr>
<tr>
<td>Seroresistant lineage averages</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.89</td>
</tr>
<tr>
<td>Moraxella catarrhalsis averages</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.91</td>
</tr>
<tr>
<td>Divergent strains</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z7574</td>
<td>XATCC23246</td>
<td>2.23</td>
<td>53.1</td>
<td>26</td>
<td>116763</td>
<td>43.60</td>
<td>Illumina MiSeq Ray (v2.3.1)</td>
<td></td>
</tr>
<tr>
<td>Z7547</td>
<td>FECCUG18198</td>
<td>2.17</td>
<td>13.3</td>
<td>46</td>
<td>87862</td>
<td>43.67</td>
<td>Illumina MiSeq Ray (v2.3.1)</td>
<td></td>
</tr>
<tr>
<td>Z7542</td>
<td>FECCUG18181</td>
<td>2.13</td>
<td>35.9</td>
<td>33</td>
<td>162472</td>
<td>43.69</td>
<td>Illumina MiSeq Ray (v2.3.1)</td>
<td></td>
</tr>
<tr>
<td>Z7546</td>
<td>FECCUG18195</td>
<td>2.12</td>
<td>3.4</td>
<td>196</td>
<td>31433</td>
<td>43.69</td>
<td>Illumina MiSeq Ray (v2.3.1)</td>
<td></td>
</tr>
<tr>
<td>Divergent strains</td>
<td>2.16</td>
<td>43.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moraxella macacae</td>
<td>2.08</td>
<td>1</td>
<td>39.00</td>
<td>454/PacBio/Illumina Ray (20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NP* = not provided, sequence obtained online from NCBI.
Table 3
Size of the Core and Supragenomes Based on Strain Number for the three individual Moraxella catarrhalis Lineages and the Combined Lineages

<table>
<thead>
<tr>
<th>Strain_added</th>
<th>Core</th>
<th>Distributed</th>
<th>Lineage Supra</th>
<th>Species Supra</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seroreistant lineage strains</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>103P14B1</td>
<td>2868</td>
<td>0</td>
<td>2868</td>
<td>2868</td>
</tr>
<tr>
<td>12P80B1</td>
<td>2564</td>
<td>266</td>
<td>2920</td>
<td>2920</td>
</tr>
<tr>
<td>43617</td>
<td>2616</td>
<td>388</td>
<td>3004</td>
<td>3004</td>
</tr>
<tr>
<td>46P47B1</td>
<td>2580</td>
<td>473</td>
<td>3053</td>
<td>3053</td>
</tr>
<tr>
<td>7169</td>
<td>2575</td>
<td>488</td>
<td>3063</td>
<td>3063</td>
</tr>
<tr>
<td>B496</td>
<td>2574</td>
<td>523</td>
<td>3097</td>
<td>3097</td>
</tr>
<tr>
<td>B502</td>
<td>2569</td>
<td>625</td>
<td>3194</td>
<td>3194</td>
</tr>
<tr>
<td>B503</td>
<td>2564</td>
<td>644</td>
<td>3208</td>
<td>3208</td>
</tr>
<tr>
<td>B507</td>
<td>2563</td>
<td>658</td>
<td>3221</td>
<td>3221</td>
</tr>
<tr>
<td>B508</td>
<td>2562</td>
<td>659</td>
<td>3221</td>
<td>3221</td>
</tr>
<tr>
<td>B583</td>
<td>2548</td>
<td>675</td>
<td>3223</td>
<td>3223</td>
</tr>
<tr>
<td>RH4</td>
<td>2545</td>
<td>690</td>
<td>3235</td>
<td>3235</td>
</tr>
<tr>
<td>BBH18</td>
<td>2541</td>
<td>699</td>
<td>3240</td>
<td>3240</td>
</tr>
<tr>
<td>Seroreistant lineage strains</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B615</td>
<td>2901</td>
<td>0</td>
<td>2901</td>
<td>3457</td>
</tr>
<tr>
<td>B616</td>
<td>2737</td>
<td>338</td>
<td>3075</td>
<td>3550</td>
</tr>
<tr>
<td>B617</td>
<td>2534</td>
<td>523</td>
<td>3157</td>
<td>3586</td>
</tr>
<tr>
<td>B618</td>
<td>2615</td>
<td>590</td>
<td>3205</td>
<td>3615</td>
</tr>
<tr>
<td>B619</td>
<td>2603</td>
<td>717</td>
<td>3320</td>
<td>3714</td>
</tr>
<tr>
<td>B620</td>
<td>2599</td>
<td>725</td>
<td>3324</td>
<td>3718</td>
</tr>
<tr>
<td>B621</td>
<td>2595</td>
<td>753</td>
<td>3348</td>
<td>3729</td>
</tr>
<tr>
<td>B622</td>
<td>2595</td>
<td>753</td>
<td>3348</td>
<td>3729</td>
</tr>
<tr>
<td>B743</td>
<td>2584</td>
<td>812</td>
<td>3396</td>
<td>3757</td>
</tr>
<tr>
<td>B744</td>
<td>2577</td>
<td>847</td>
<td>3424</td>
<td>3767</td>
</tr>
<tr>
<td>B745</td>
<td>2572</td>
<td>865</td>
<td>3437</td>
<td>3780</td>
</tr>
<tr>
<td>B746</td>
<td>2569</td>
<td>883</td>
<td>3452</td>
<td>3791</td>
</tr>
<tr>
<td>B747</td>
<td>2565</td>
<td>898</td>
<td>3463</td>
<td>3801</td>
</tr>
<tr>
<td>B748</td>
<td>2560</td>
<td>952</td>
<td>3512</td>
<td>3845</td>
</tr>
<tr>
<td>B749</td>
<td>2551</td>
<td>963</td>
<td>3514</td>
<td>3845</td>
</tr>
<tr>
<td>B750</td>
<td>2550</td>
<td>964</td>
<td>3514</td>
<td>3845</td>
</tr>
<tr>
<td>MX1</td>
<td>2509</td>
<td>1005</td>
<td>3514</td>
<td>3845</td>
</tr>
<tr>
<td>S11</td>
<td>2464</td>
<td>1060</td>
<td>3524</td>
<td>3854</td>
</tr>
<tr>
<td>SR/SS</td>
<td>2318</td>
<td>1536</td>
<td>NA</td>
<td>3854</td>
</tr>
<tr>
<td>Divergent strains</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z7542</td>
<td>4106</td>
<td>0</td>
<td>4106</td>
<td>5518</td>
</tr>
<tr>
<td>Z7546</td>
<td>3938</td>
<td>168</td>
<td>4274</td>
<td>5518</td>
</tr>
<tr>
<td>Z7547</td>
<td>3780</td>
<td>179</td>
<td>4284</td>
<td>5518</td>
</tr>
<tr>
<td>Z7574</td>
<td>3632</td>
<td>180</td>
<td>4285</td>
<td>5684</td>
</tr>
<tr>
<td>SS/SS/Divergent</td>
<td>2041</td>
<td>3643</td>
<td>NA</td>
<td>5684</td>
</tr>
</tbody>
</table>

**Note:** Boldface indicates lineage, or lineage combination values.

In Table 3, we present the size of the core and supragenomes based on strain number for the three individual Moraxella catarrhalis lineages and the combined lineages. The table shows the number of genes/genome type for each strain added, including core, distributed, lineage supra, and species supra.

Phylogeny and Evolution of the Moraxella Catarrhalis Species

We performed three different neighbor-net analyses for the 35 strains (SR, SS, and divergent lineages) using the SplitsTree4 program (Huson and Bryant 2006) in which we used as input 1) allelic data for all of the core genes (fig. 3A); 2) gene possession data for all of the distributed genes (fig. 3B); and 3) a multiple sequence alignment of a concatenation of the LCBs (physical core regions irrespective of gene annotations) as defined by a default MAUVE (Darling et al. 2004) alignment of all 31 M. catarrhalis genomes (fig. 3C). Finally, we also performed a neighbor-net analysis of the concatenated collinear blocks between all M. catarrhalis strains and the M. macacae genome (representing an outgroup) (fig. 3D). Each of these independent methods demonstrated a clear separation between the SS and SR lineages, but the two SS ribotypes were randomly distributed within the larger SS clade and did not cluster together. It is also apparent that the four divergent strains cluster together, but that they are vastly more different from both the SS and SR lineages than the SS and SR lineages are from each other. All three comparative measures suggest that the divergent strains are slightly closer to the SS strains than the SR strains, but the total distance to the SS strains is >90% of the distance to the SR strains. The reticulations (webbing) present within these diagrams are indicative of ambiguous ancestry, but given that M. catarrhalis is naturally highly competent and transformable these findings are suggestive of HGT events as they represent alternative phylogenetic branching; however, it is also possible that they represent parallel mutations, or the loss of ancestral copies in one or the other lineages. These reticulations are present both within the individual clades and are combined, there was a substantial increase in diversity ($X^2 = 413 \pm 40$); but the diversity reached its zenith with the addition of the divergent strains ($X^2 = 821 \pm 809$) (table 4). This trend of increasing diversity is also reflected in the standard deviation values for the gene similarity metrics.

Analyses of the various distributed gene frequency classes of the combined SR and SS supragenome revealed small, but easily identifiable, increases in the number of distributed genes that are present in 13 and 18 genomes, relative to the number of distributed genes in the other ($n \neq 13$; $n \neq 18$) intermediate gene frequency classes. These findings correspond with the exact numbers of SR and SS strains in the analysis, and are indicative of the small numbers of genes that are exclusively core to each of these lineages (fig. 2A). As expected, a gene-by-gene analysis using our standard clustering criteria (70% homology over 70% of the alignment) (Hogg et al. 2007) showed that in spite of the high degree of similarity between the SR and SS core genomes, there are 33 SR core-specific genes that are not found in any of the SS strains (table 5). Conversely, there are 49 SS lineage core genes that are not found in any of the SR strains. A similar analysis demonstrated a large peak at the gene cluster class size of four when the divergent strains were included in the analysis, indicating that the four divergent strains harbor a unique set of core genes that are not present in the SR and SS lineages (fig. 2B). Annotations for all M. catarrhalis genes for all of the sequenced strains are available in GenBank (table 1).
lineages, and are also present between the lineages. Such findings are consistent with interlineage gene flow, but could also indicate parallel mutations or loss of ancestral genes. In particular, it is clear that there is substantial distributed gene flow between the SR and SS lineages, indicating that they are not currently reproductively isolated. The decreased number of reticulations in the concatenated core region tree is a direct result of the selection process for inclusion of regions in the analysis as only regions that are core across all strains were included. The concatenation network in which we included *M. macacae* as an out group showed that it was approximately equidistant from all three lineages.

We then evaluated the various groupings for evidence of diversity generation by determining both the intra- and interlineage levels of HGT, and by determining the extent of allelic diversity as measured by SNPs of core genes within and between each group as a measure of their relative phylogenetic ages. These last two measures were designed to determine if there were different evolutionary mechanisms acting on the two main lineages as has been previously suggested (Wirth et al. 2007). These investigators speculated that the SR and SS lineages evolved primarily by different mechanisms (Wirth et al. 2007). However, we find no support for this conjecture as it is clear that HGT of distributed genes is just as prevalent among the SS lineage as the SR lineage, and that...
This distributed gene which has two alleles (bro1 and bro2), either of which were found inserted between the amidotransferase genes of a given strain, was found in all but one of the SR lineage strains, and in slightly more than half of the SS strains; it was not identified in any of the divergent strains. Independent phylogenetic trees were built using phylm from mauve alignments containing the amidotransferase A and bro genes, respectively (fig. 6A and B). Examination of the amidotransferase A tree establishes that this gene is frequently moving via HGT both within each of the SR and SS clades, but also between the clades as well. We see this gene tree structure, which is very different from a consensus tree for the species as a whole, presumably because of the positive selective pressure produced by acquisition of the adjacent bro gene which provides for high-level penicillin resistance. Examination of the data suggests that the SS strains F18 and MX1 acquired their amidotransferase and bro1 genes via HGT from an SR strain within/related to the BC7-containing group of SR strains. Similarly, it is likely that the O35E and BC1 SR strains acquired their amidotransferase A and bro1 genes via HGT from an SS strain within/related to the F-21/F24 containing group. Within the BC7 group of SR strains there has also been loss of the bro1 gene (4P4681) followed by acquisition of the bro2 allele in several strains (B8H18, 12P80B1, and 43617). The β-lactamase gene itself was not evaluated in the orderedPainting analysis because it is a distributed gene, albeit a highly prevalent one.

The independent evolution of the SS and SR lineages into distinct clades with neither appearing to be ancestral to the other raises two questions: When did the split occur that resulted in their isolation, and when were they reunited? Toward this end we attempted to perform an evolutionary divergence analysis with BEAST (Drummond and Rambaut 2007); however, due to our having only a single unverifiable calibration point extreme care must be used in interpreting these data. The BEAST results suggest an earlier split between the M. macacae and M. catarhalis populations at 37 Ma with an effective sample size (ESS) score of 118, and a split between the SS and SR M. catarhalis clades at 0.265 Ma with an ESS score of 3 after 10 million iterations. It is important to note that an ESS score of less than 100 is considered not significant; however, without intermediate time points (i.e., additional phylogenetic branches within Moraxella) it is not possible to further refine this estimate.

### Virulence Factor Distribution

To date, numerous *M. catarhalis* virulence factors have been described that play a role in immune evasion, cellular adherence, biofilm formation, and survival during iron starvation, which have been reviewed (de Vries et al. 2009; Su et al. 2012). Previous evaluation of virulence factor distribution in 12 SR lineage isolates (Davie et al. 2011) demonstrated that, with a few exceptions, all known virulence factors were
present in each SR lineage strain. In this study, we found 33 gene clusters that were core to the SR lineage and missing from all of the SS strains (table 5). None of these gene clusters are obvious virulence factors; however, the majority of these clusters that are unique to the SR lineage either are predicted to encode hypothetical proteins of unknown function or are under-annotated, suggesting that there remains much important biology to be developed to understand the observed phenotypic differences. Within the small group of SR-specific genes, almost all of those that were annotated were associated with phosphate metabolism, including at least ten genes with putative annotations as phosphodiesterases, phosphatases, phosphate permeases, and adenosine triphosphate (ATP) transporters clustered together on the chromosome and likely assembled into one or more tightly linked operons. In addition, there were also a number of orphan

---

**Fig. 2.**—Number and distribution of gene clusters as a function of frequency within the species, and as the percent similarity requirement increases. x-axis = number of genomes containing a given gene cluster; y-axis = number of gene clusters; z-axis = percent identity to be grouped in a cluster. (A) Analysis of the combined SR and SS lineages. Note the small increases in the gene frequency classes at 13 and 18 which correspond to the numbers of SR and SS strains analyzed; these represent genes which are uniquely core to each lineage; (B) analysis of the combined SR, SS, and divergent lineages. Note the very large increase in the gene frequency class at four genomes corresponding to number of divergent strains in the analysis; this spike represents genes that are core to the divergent lineage which are not present in either the SR or SS lineages.
The known virulence factor genes were found also, somewhat surprisingly, to be nearly universally conserved across the panel of sequenced SS strains (supplementary table S5, Supplementary Material online). One of the few exceptions was the gene encoding the Opa-like protein A (OpaA) which was absent from strain ATCC 43617, confirming previous findings (Brooks et al. 2007). The prevalence of genes encoding the M. catarrhalis filamentous hemagglutinin-like proteins (Mha/Mch) (Baider et al. 2007; Plamondon et al. 2007) could not be accurately analyzed due to high sequence similarity and inconsistent assembly of whole-genome sequence data, as previously pointed out (Davie et al. 2011). All of the strains contained the uspA2/A2H gene. The uspA1 gene was detected in all SR isolates and in 16 of the 18 SS isolates, with isolates N4 and N15 being uspA1 negative, confirming previous PCR-screening results by Bootsma et al. (1996). These results are comparable with those reported in a publication by Verhaegh et al. (2011) who showed that in a panel of 112 M. catarrhalis isolates (n = 22 SR lineage, n = 90 SS lineage based on ribotyping), 97% were uspA1 PCR positive, and all strains were PCR positive for either uspA2 or uspA2H. In that study, mid/haq was only detected by PCR in 80% of isolates, while it was present in 100% of the isolates in our study.

Interestingly, the presence of the β-lactamase gene bro-1/2 differed between the lineages as out of the 18 sequenced SS strains, 8 did not harbor the β-lactamase gene, in contrast to only two bro-1/2 negative SR isolates. In silico typing of LOS gene clusters (Edwards et al. 2005) demonstrated that most isolates harbor the type A gene cluster (24/31), although the prevalence was higher in the sequenced SS lineage isolates (17/18) than in the SR lineage (7/13), which is in line with previous observations (Verhaegh et al. 2008). The presence of lgt4, encoding a predicted N-acetylglucosamine transferase, is characteristic for LOS type A and C strains (Peak et al. 2007). In our panel of sequenced isolates, an LOS type B gene cluster lacking lgt4 was only found in four SS lineage isolates and not in any SS lineage isolates. Furthermore, the LOS type C gene cluster was detected in 1 and 2 SS and SR lineage isolates, respectively.

### Table 5

<table>
<thead>
<tr>
<th>BBH18 Genes</th>
<th>BBH18_annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCR_0971</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>MCR_0970</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>MCR_0969</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>MCR_0968</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>MCR_0646</td>
<td>IS200 family transposase</td>
</tr>
<tr>
<td>MCR_0640</td>
<td>CRISPR-associated protein NE0113 (Cas_NE0113); family protein</td>
</tr>
<tr>
<td>MCR_0623</td>
<td>Acid phosphatase autotransporter</td>
</tr>
<tr>
<td>MCR_0619</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>MCR_1427</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>MCR_0008</td>
<td>Glycerolphosphoryl diester phosphodiesterase</td>
</tr>
<tr>
<td>MCR_1871</td>
<td>Putative membrane protein</td>
</tr>
<tr>
<td>MCR_1721</td>
<td>HAD-supersfamily subfamily IB hydrolase</td>
</tr>
<tr>
<td>MCR_1723</td>
<td>Polyamine ABC transporter permease protein</td>
</tr>
<tr>
<td>MCR_1725</td>
<td>Type I phosphodiesterase/nucleotide; pyrophosphatase</td>
</tr>
<tr>
<td>MCR_1726</td>
<td>Extracellular solute-binding protein family 1; protein</td>
</tr>
<tr>
<td>MCR_1731</td>
<td>Phosphate ABC transporter ATPase subunit PstB</td>
</tr>
<tr>
<td>MCR_1732</td>
<td>Phosphate ABC transporter permease protein PstA</td>
</tr>
<tr>
<td>MCR_1733</td>
<td>Phosphate ABC transporter permease protein PstC</td>
</tr>
<tr>
<td>MCR_1734</td>
<td>Phosphate ABC transporter substrate binding; protein PstS</td>
</tr>
<tr>
<td>MCR_0920</td>
<td>Conserved hypothetical protein</td>
</tr>
<tr>
<td>MCR_1199</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>MCR_1722</td>
<td>Polyamine ABC transporter ATPase subunit</td>
</tr>
<tr>
<td>MCR_1724</td>
<td>Polyamine ABC transporter permease protein</td>
</tr>
<tr>
<td>MCR_1013</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>MCR_1123</td>
<td>Conserved hypothetical protein</td>
</tr>
<tr>
<td>MCR_0622</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>MCR_0811</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>Unannotated</td>
<td>Annotions</td>
</tr>
<tr>
<td>BBH18 Genes</td>
<td>BBH18_annotations</td>
</tr>
<tr>
<td>103P14B1_285</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>12P80B1_301</td>
<td>Serine protease</td>
</tr>
<tr>
<td>12P80B1_1567</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>B406_1518</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>B583_283</td>
<td>Hypothetical protein</td>
</tr>
</tbody>
</table>

Note.—BBH18 genes refer to SR core genes that were annotated in the SR strain BBH18 [13]. This genome was chosen as it has been closed. Unannotated BBH18 genes refer to SR core genes that were annotated in one of the other SR strains, but not in BBH18; however, all SR strains including BBH18 were determined by FASTA alignment [16,44] to contain these genes.

Phenotype–Genotype Association

To enable phenotype–genotype association analyses for key virulence determinants in M. catarrhalis, that is, complement resistance and cellular adherence, the SS and SR strains used in these analyses were subjected to phenotyping with respect to their ability to survive in 40% NHS and adherence to respiratory tract epithelial cells. These studies demonstrated that all but one (ATCC 43617) of the strains previously characterized as SS were highly sensitive to killing by NHS (fig. 7A). Similarly, all but two (C072 and 7169) of the SR lineage strains showed efficient adherence to A549 type II alveolar cells and Detroit 562 pharyngeal epithelial cells, whereas all but one (N1) of the SS strains adhered poorly to these human cells types (fig. 7B). All these phenotypic findings were statistically significant. Next we analyzed the association of M. catarrhalis virulence phenotypes (fig. 7A and B) with genotypes using PhenoLink.
However, given the clear separation of the SR and SS lineages and the ability of the SR stains to resist complement-mediated killing, the value of this analysis was limited, yielding primarily the lineage-specific genes identified above. Moreover, manual comparison of the SR lineage isolates with strain ATCC 43617, the only complement-sensitive SR lineage isolate in our panel, showed that there are no genes that were absent only from the ATCC 43617 genome, suggesting this defect is allelic in nature. Similar observations were made for association with A549 adherence phenotypes, as all SS lineage isolates were found to attach to A549 cells with low efficiency. Comparing the low binding SR isolates (7169 and C072) with the intermediate and/or high binding SR isolates did not reveal any genes that were lacking specifically from these two isolates. Association analyses for the Detroit 562 adherence phenotype showed that two

**FIG. 3.**—NeighborNet analyses performed using SplitsTree4. Branch length is proportional to the degree of variation between and among strains. Reticulations (webbing) are indicative of HGT as they represent ambiguities with respect to lineal descent. Blue boxes indicate SS strains; red boxes indicate SR strains; and yellow boxes indicate divergent strains. (A) Network prepared using core gene allelic data; (B) network prepared using presence/absence of distributed gene clusters; (C) network prepared using concatenated core sequences independent of annotation; (D) network prepared using core sequence alignment shows phylogenetic distance between *Moraxella catarrhalis* and *Moraxella macacae*. The latter was used as outgroup to root the tree and to look for any ancestral relationships within the *M. catarrhalis* species. Note that the greatest degree of HGT both within and between lineages is observed with the distributed gene analyses; however, it is also evident that the core genes and regions are exchanged as well.

(Bayjanov et al. 2012).
genes \((MCR_{1831}\text{ and }MCR_{1286},\text{ encoding a type I restriction modification DNA specificity protein and a membrane protein-like protein, respectively})\) were present in the high binding SS isolate N1 and absent from all other SS lineage isolates; however, these genes were also found to be present in the low binding SR lineage isolates 103P14B1, 12P80B1, and BC7.

**Discussion**

The relative overall average pairwise gene possession differences among the SR lineage strains (Davie et al. 2011) is only 232 ± 55, which is substantially less than has been observed for two other supragenomically characterized nasopharyngeal pathogens of similar genomic size—H. influenzae (395 ± 4) (Hogg et al. 2007) and S. pneumoniae (407 ± 91) (Hiller et al. 2007). With the inclusion of the SS lineage strains in this study, the species-level pairwise gene possession differences (412 ± 140) very nearly approximate those of the NTHi and pneumococcus. The somewhat larger SD value is attributable to the \(M.\ catarrhalis\) species being composed of two distinct lineages such that all cross-lineage pairwise comparisons result in very much higher numbers of differences than do any of the within lineage comparisons.
Moreover, the supragenomes for the SR, SS, and combined SS/SR *M. catarrhalis* groupings based on the FSM predictions are very close in size. This is an indication that there is very little divergence between the two lineages and is highly supportive of a single species model. If we had observed a dramatic decrease in the size of the *M. catarrhalis* SR core genome upon the addition of the SS lineage strains, as was seen with the combining of multiple *Gardnerella vaginalis* clades (Ahmed et al. 2012), that would have been supportive for a species-level taxonomic split. In spite of the relative similarity between the SR and SS clades, they each possess a small clade-specific core genome. These genes are, respectively, excellent candidates for virulence analyses and suppressors of pathogenicity. It is tempting to hypothesize that the cluster of genes associated with phosphate metabolism (table 5) that are core to the SR lineage may be involved in a nucleotide second messenger system that controls virulence gene expression (D’Argenio and Miller 2004). The fact that many traditional virulence genes were highly conserved across the SS and SR lineages indicates that individual gene presence is not predictive of virulence, which suggests that it may be gene combinations, differences in expression, or environmental conditions that in concert control differences with regard to virulence potential. In support of this last hypothesis it has been shown that host factors play a role in determining whether/when given *M. catarrhalis* strains are pathogenic or not as previous studies have shown that under proinflammatory conditions that host tissues upregulate CEACAMS which then provide for better binding by SS strains (Hill and Virji 2003; Hill et al. 2012).

The vast majority of *M. catarrhalis* strains that are recovered from respiratory tract infections belong to the SR lineage and it has been noted for the past quarter century that almost all such isolates are penicillin resistant due to the possession of a β-lactamase gene (*bro*) (Bluestone et al. 1992; Johnson et al. 2003). Using an unbiased screen for identifying core chromosomal sites undergoing high rates of HGT, we identified an amidotransferase operon. Examination of this locus revealed that it is the sole chromosomal site for insertion of the distributed *bro* gene for both the SR and SS lineages. The high rate of recombination at this site combined with the very limited heterogeneity of the *bro* genes, in contrast to the high degree of allelic variability among the adjacent amidotransferase A genes, is supportive of a recent acquisition history for the *bro* gene followed by rapid spread throughout the *M. catarrhalis* species confirming earlier findings (Bootsma et al. 1996, 1999, 2000b).

In the current analyses, we also included a group (n = 4) of highly divergent strains that had previously been characterized as distantly related to *M. catarrhalis* (With et al. 2007). The inclusion of these strains essentially doubled the average number of pairwise gene possession differences (821 ± 809) when compared with the SR/SS species-level view. What was also striking about the supragenome comparisons when these divergent strains were added was that we obtained some negative pairwise comparison scores (table 4) which means that some strains shared less than 50% of their genes. In our analyses of the supragenomes of some two dozen bacterial species, the only other time that we obtained negative-value
comparison scores was with Gardnerella vaginalis (Ahmed et al. 2012) which turned out to actually be a genus-level classification composed of multiple species.

As with the Gardnerella work we now propose that the divergent strains belong in a separate taxonomic grouping. Although the definition of a bacterial species is currently under critical review by the scientific community, our suggestion is supported by multiple types of analyses and metrics that are both broadly employed, and currently viewed as being the current “best estimate” used to define a split between different strains and different species. The currently accepted level of similarity to be included in the same species with regard to nucleotide identity is an ANI of 95–96%, and a corresponding tetranucleotide identity of >99% (Konstantinidis and Tiedje 2005; Goris et al. 2007; Richter and Rosello-Mora 2009; Chan et al. 2012). Using these criteria, these data are supportive of the SR and SS strains belonging to the same species. However, the average pairwise comparisons between SS/ SR strains and the divergent strains gave ANI values of 89.01% and tetranucleotide values of 95.57%, both far outside the accepted range for a species.

Similarly, each species has a very narrowly defined percent GC content which always varies by less than 1%, and usually by less than 0.5%, among the component strains. All of the SS and SR lineage strains have GC contents within 0.3% of one another, again supporting their classification as a single species. As with the ANI values, the four divergent strains cluster together very closely among themselves, but their GC content of 43.66% differs by more than 2% from the SR/SS lineages, again clearly indicating that they are sui generis. Finally, analysis of the effect on the core genome size by combining the various lineages is also very revealing with regard to identifying natural genomic proclivities. Riley and Lizotte-Waniewski (2009), Donati et al. (2010) have proposed using the core genome as a test of whether individual strains should be grouped within the same species; they showed that even a single Streptococcus mitis strain’s genome added to the S. pneumoniae supragenome resulted in a precipitous drop in the combined core genome size. Thus, we have proposed a bacterial species definition based on core genome relatedness such that any strain that results in less than a 10% decrease in the size of an established (at or near its asymptote) core genome belongs within that species and conversely those that result in a greater than 10% decrease belong in a different species (Nistico et al. 2014). Although many of the pathogenic enterobacteriaceae species such as Escherichia coli have very large supragenomes (Ahmed et al. 2012), comparison across clades shows little variability among core genomes. In contrast, the addition of just a single member of the M. catarrhalis divergent strain group to the combined SS/ SR core genome results in decrease in size of greater than 11%. Collectively, these data are supportive of the current taxonomy in which the SR and SS lineages form a single M. catarrhalis species, but suggest that the four divergent strains may be considered a distinct genospecies.

The finding that the M. macacae is approximately equidistantly related to the two (SR and SS) M. catarrhalis lineages is suggestive that neither of the lineages are directly ancestral to the other. However, the independent evolution of the SS and SR lineages raises two questions: When did the split occur that resulted in their isolation, and when were they reunited? Based on their possession of nearly identical core genomes and the amount of horizontal gene flow between the two lineages, they are clearly still a single species by any criteria; however, the low-resolution BEAST analysis we performed suggested a split at ~265,000 BP. It is interesting to speculate that one of the lineages may have evolved independently in the Neanderthal population which diverged from the ancestors of modern humans ~ 370 Ma (Noonan et al. 2006) and was then subsequently reintroduced into anatomically modern humans, where the other lineage evolved, during the short period of geographic and social overlap of these
**Fig. 6.**—(A) Phylogenetic tree built using phyml from a mauve alignment of the amidotransferase subunit A genes. SR strains are shown in red typeface; SS strains are shown in blue typeface; and the divergent strains are shown in yellow typeface. “+” indicates the presence of the β-lactamase gene in the strain, and “-” indicates a lack of the β-lactamase gene in the tree. Bro1 and Bro2 following the + after the strain name refers to which of the two β-lactamase alleles are present in the strain; and (B) phylogenetic tree built using phyml from a mauve alignment of the bro gene (bro); note that there are only two alleles of the bro gene distinguished by five-point mutations of which all but one are synonymous. The bro2 containing strains are on the left, and the bro1 containing strains are on the right. SR strains are shown in red typeface; SS strains are shown in blue typeface.
distinct people (Green et al. 2010) in Europe approximately 30–45,000 years before the present following the latter's migration out of Africa. Although the estimated separation dates are somewhat different, they are still supportive of the general hypothesis of short-term independent lineage evolution as it is very difficult to perfectly synchronize evolutionary clock rates across species particularly those as divergent as microbe and man. However, these data could also support other early human population splits and reunions.

Summary

The analyses presented herein support the hypothesis that the *M. catarrhalis* species consists of two distinct lineages, the SR and the SS, with separate evolutionary histories and small, but distinct, lineage-specific core genomes; however, it is also clear that the two clades have been reunited and now share a single environmental niche and regularly exchange DNA via HGT. Thus, the questions of how and when the lineages diverged and then at some later point reconverged need to be further addressed. It is also clear that the divergent strains form a distinct genospecies separate from *M. catarrhalis*.

Supplementary Material

Supplementary tables S1–S5 and figure S1 are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by Allegheny Singer Research Institute, Allegheny General Hospital, and Drexel University College of Medicine with grant funding from the United States National Institutes of Health grants DC05659, AI080935, and DC02148, to G.D.E., and from a Vienna Spot of Excellence grant (ID337956). The authors thank Mark Achtman for providing access to the DNA for the divergent strains and for facilitating their whole-genome sequencing; Joshua Chang Mell for fruitful discussions; and Ms Mary O’Toole and Ms Carol Hope for help in the preparation and submission of this manuscript.

Literature Cited


**Associate editor:** Tal Dagan